A CMA-ES for Mixed-Integer Nonlinear Optimization

Nikolaus Hansen

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A CMA-ES for Mixed-Integer Nonlinear Optimization

Nikolaus Hansen

N° 7751

October 2011

Thème COG

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--7751--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SACLAY - ÎLE-DE-FRANCE

A CMA-ES for Mixed-Integer Nonlinear Optimization

Nikolaus Hansen*

Thème COG — Systèmes cognitifs
Équipes-Projets Adaptive Combinatorial Search et TAO

Rapport de recherche n° 7751 — October 2011 — 13 pages

Abstract: We propose a modification of CMA-ES for the application to mixed-integer problems. The modification is comparatively small. First, integer variables with too a small variation undergo an additional integer mutation. This mutation is also used for updating the distribution mean but disregarded in the update of covariance matrix and step-size. Second, integer variables with too a small variation are disregarded in the global step-size update alltogether. This prevents random fluctuations of the step-size.

Key-words: optimization, evolutionary algorithms, CMA-ES, mixed integer optimization, black-box optimization

* NH is with the TAO Team of INRIA Saclay–Île-de-France at the LRI, Université-Paris Sud, 91405 Orsay cedex, France

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université
4, rue Jacques Monod, 91893 ORSAY Cedex
Téléphone : +33 1 72 92 59 00

No French Title

Résumé : Pas de résumé

Mots-clés : Pas de motclef

1 Introduction

The *covariance matrix adaptation evolution strategy* (CMA-ES) [7, 8, 5, 4, 2, 6] is a stochastic, population-based search method in continuous search spaces, aiming at minimizing an objective function $f : \mathbb{R}^n \to \mathbb{R}, \boldsymbol{x} \mapsto f(\boldsymbol{x})$ in a black-box scenario. In this report we describe two small modifications of CMA-ES that are sometimes useful and sometimes even essential in order to apply the method to mixed-integer problems.

We assume that for some of the components of the vector $\boldsymbol{x} \in \mathbb{R}^n$ only integers are feasible values for evaluating f. This is easy to ensure: the components can be rounded or truncate when evaluating f (both should be equivalent, because a solver should be translation invariant). In general, this is a feasible way to handle variables with any (preferably equi-distant) granularity. From the solvers view point, the function becomes in some coordinates piece-wise constant. Consequently, this report can be interpreted in describing the pitfalls and their remedies of solving continuous but partly variable-wise constant functions with CMA-ES.

The problem There occurs one main problem when solving piecewise constant functions with CMA-ES. When the sample variance of a component is small, *all* points might be sampled into the same constant section for this component. More precisely this happens, when the standard deviation is much smaller than the stair-width of the granularity (which could be assumed w.l.o.g. to be one). In this case, selection in this component becomes "arbitrary". There is no particular reason that the standard deviation of this component will increase again, because the parameter chances in the CMA-ES are unbiased. Even worse, because the overall step-size usually decreases over time, there is a good chance that the component under consideration will never be tested with a different value again for the remainder of the run.

The solution This problem is solved by introducing an integer mutation for components where the variance appears to be small. The integer mutation is disregarded for the update of any state variable but the mean. Additionally, for step-size adaptation, components where the integer mutation is dominating are out-masked. The details are given in the next section.

2 The $(\mu/\mu_\mathrm{w}, \lambda)$-CMA-ES

In the standard $(\mu/\mu_\mathrm{w}, \lambda)$-CMA-ES, in each iteration step k, λ new solutions $\boldsymbol{x}_i \in \mathbb{R}^n$, $i = 1, \ldots, \lambda$, are generated by sampling a multi-variate normal distribution, $\mathcal{N}(\mathbf{0}, \boldsymbol{C}_k)$, with mean $\mathbf{0}$ and $n \times n$ covariance matrix $\boldsymbol{C}_k$, see Eq. (1). The μ best solutions are selected to update the distribution parameters for the next iteration $k + 1$. The complete algorithm is depicted in Table 1.

The stair-width diagonal matrix $\boldsymbol{S}^{\mathrm{int}}$ has zero entries for continuous variables. For variables with non-zero granularity the value $s[x/s]$ is used for evaluating the objective function, where x is the variable value and s the respective diagonal entry of $\boldsymbol{S}^{\mathrm{int}}$. The iteration index is $k = 0, 1, 2, \ldots,$ where $\boldsymbol{p}^\sigma_{k=0} = \boldsymbol{p}^\mathrm{c}_{k=0} = \mathbf{0}$ and $\boldsymbol{C}_{k=0} = \boldsymbol{I}$. Here, $\boldsymbol{x}_{i:\lambda}$ is the i-th best of the solu-

Table 1: Update equations at iteration k for the state variables in the $(\mu/\mu_\mathrm{w}, \lambda)$-CMA-ES allowing for variable granularities according to $\boldsymbol{S}^\mathrm{int}$. Shaded areas are modifications of the original algorithm, where $\boldsymbol{s}_{i:\lambda} = \boldsymbol{S}^\mathrm{int} R^\mathrm{int}_{i:\lambda}/\sigma_k$, but (5) is unmodified, because $\alpha = 0$ by default. R^int_i typically has a single component set to ± 1 and $\boldsymbol{I}^\sigma_k$ is a masking matrix. In contrast to the original algorithm, $\boldsymbol{m}_{k+1} - \boldsymbol{m}_k \neq \sigma_k \sum_i^\mu w_i\, \boldsymbol{y}_{i:\lambda}$, which led to a different way to write (3) and (4)

Given $k \in \mathbb{N}$, $\boldsymbol{m}_k \in \mathbb{R}^n$, $\sigma_k \in \mathbb{R}_+$, $\boldsymbol{C}_k \in \mathbb{R}^{n\times n}$ positive definite, $\boldsymbol{p}^\sigma_k \in \mathbb{R}^n$, and $\boldsymbol{p}^\mathrm{c}_k \in \mathbb{R}^n$

$$\boldsymbol{x}_i \sim \boldsymbol{m}_k + \sigma_k \boldsymbol{y}_i + \boldsymbol{S}^\mathrm{int} R^\mathrm{int}_i \qquad \text{where } \boldsymbol{y}_i \sim \mathcal{N}_i(\boldsymbol{0}, \boldsymbol{C}_k) \text{ for } i = 1, \ldots, \lambda \tag{1}$$

$$\boldsymbol{m}_{k+1} = \sum_{i=1}^{\mu} w_i\, \boldsymbol{x}_{i:\lambda} \qquad \text{where } f(\boldsymbol{x}_{1:\lambda}) \leq \cdots \leq f(\boldsymbol{x}_{\mu:\lambda}) \leq f(\boldsymbol{x}_{\mu+1:\lambda}) \ldots \tag{2}$$

$$\boldsymbol{p}^\sigma_{k+1} = (1 - c_\sigma)\, \boldsymbol{p}^\sigma_k + \sqrt{c_\sigma(2 - c_\sigma)\mu_\mathrm{w}}\; \boldsymbol{C}_k{}^{-\frac{1}{2}} \sum_{i=1}^{\mu} w_i\, \boldsymbol{y}_{i:\lambda} \tag{3}$$

$$\boldsymbol{p}^\mathrm{c}_{k+1} = (1 - c_\mathrm{c})\, \boldsymbol{p}^\mathrm{c}_k + h^\sigma_k \sqrt{c_\mathrm{c}(2 - c_\mathrm{c})\mu_\mathrm{w}} \sum_{i=1}^{\mu} w_i\, \boldsymbol{y}_{i:\lambda} \tag{4}$$

$$\begin{aligned}\boldsymbol{C}_{k+1} = {} & (1 - c_1 - c_\mu + (1 - h^\sigma_k)\, c_1 c_\mathrm{c}(2 - c_\mathrm{c}))\, \boldsymbol{C}_k \ldots \\ & + c_1\, \boldsymbol{p}^\mathrm{c}_{k+1} \boldsymbol{p}^\mathrm{c}_{k+1}{}' + c_\mu \sum_{i=1}^{\mu} w_i (\boldsymbol{y}_{i:\lambda} + \alpha \boldsymbol{s}_{i:\lambda})(\boldsymbol{y}_{i:\lambda} + \alpha \boldsymbol{s}_{i:\lambda})'\end{aligned} \tag{5}$$

$$\sigma_{k+1} = \sigma_k \times \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|\, \boldsymbol{I}^\sigma_k\, \boldsymbol{p}^\sigma_{k+1} \|}{\mathsf{E}\|\mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}^\sigma_k\right)\|} - 1\right)\right) \quad \text{if } \boldsymbol{I}^\sigma_k \text{ has non-zero components, } \sigma_{k+1} = \sigma_k \text{ otherwise} \tag{6}$$

tions $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\lambda$ and $i : \lambda$ denotes the respective index and $h^\sigma_k = 1$ if $\|\boldsymbol{p}^\sigma_{k+1}\| < \sqrt{1 - (1 - c_\sigma)^{2(t+1)}} \left(1.4 + \frac{2}{n+1}\right) \mathsf{E}\|\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})\|$ and zero otherwise.

The masking matrix $\boldsymbol{I}^\mathrm{R}_k$ is a diagonal matrix that signifies the components, where an integer mutation will be conducted. Diagonal elements are one where for the respective diagonal element $2\sigma_k \boldsymbol{C}_k{}^{\frac{1}{2}} < \boldsymbol{S}^\mathrm{int}$, and zero otherwise. The λ^int_k determines the number of candidate solutions that will (probably) receive an integer mutation. The R^int_i is the integer mutation, for individuals $i = 1, \ldots, \lambda^\mathrm{int}_k$. We have $R^\mathrm{int}_i = \boldsymbol{I}^{\pm 1}_i (R'_i + R''_i) \in \mathbb{R}^n$, where the sign-switching matrix $\boldsymbol{I}^{\pm 1}_i$ is a diagonal matrix with diagonal elements being ± 1 each with probablity $1/2$, independently for all components and all i. The R'_i has exactly one of its components, those unmasked by $\boldsymbol{I}^\mathrm{R}_k$, set to one, such that $\boldsymbol{I}^\mathrm{R}_k R'_i \neq \boldsymbol{0}$. Additionally the R'_i are dependent in that the number of mutations for each coordinate differs at most by one (e.g. each coordinate has received zero or one mutation). In R''_i, the same components have a geometric distribution with $p = P(X = 0) = 0.7^{1/|\boldsymbol{I}^\mathrm{R}_k|}$, which means $R''_i = \boldsymbol{I}^\mathrm{R}_k R''_i$ is usually the zero vector.

Additionally, if $\lambda^\mathrm{int}_k > 0$, we set $R^\mathrm{int}_\lambda = \lfloor \boldsymbol{S}^{\mathrm{int}^{-1}} \boldsymbol{x}^{k-1}_{1:\lambda} \rfloor - \lfloor \boldsymbol{S}^{\mathrm{int}^{-1}} \boldsymbol{m}_k \rfloor$, where $\boldsymbol{S}^{\mathrm{int}^{-1}}$ equals $\boldsymbol{S}^\mathrm{int}$ element-wise inverted in all non-zero elements of $\boldsymbol{S}^\mathrm{int}$. This means, the last candidate solution gets its integer values from the previous best solution. Otherwise, R^int_i equals the zero vector, for $i = 1, \ldots, \lambda$.

We set $\lambda_k^{\text{int}} = \min(\lambda/10 + |\boldsymbol{I}_k^{\text{R}}| + 1, \lfloor\lambda/2\rfloor - 1))$ if $0 < |\boldsymbol{I}_k^{\text{R}}| < n$ and $\lambda_k^{\text{int}} = 0; \lfloor\lambda/2\rfloor$ if $|\boldsymbol{I}_k^{\text{R}}| = 0; n$.

Similar to $\boldsymbol{I}_k^{\text{R}}$, the masking matrix $\boldsymbol{I}_k^{\sigma}$ is the n-dimensional identity matrix with the j-th element set to zero if the j-th diagonal element of $\sigma_k \, \boldsymbol{C}_k^{\frac{1}{2}} / \sqrt{c_\sigma}$ is smaller than the j-th diagonal element of $0.2\,\boldsymbol{S}^{\text{int}}$.

We set $\alpha = 0$, because we found so far no empirical evidence for using the integer variations in the covariance matrix update with $\alpha = 1$ being better. The integer mutations are carried out independently of the covariance matrix in any case. Further symbols and constants are given below.

The integer version deviates in Equations (1) and (6) from the original version. For $\boldsymbol{S}^{\text{int}} = \mathbf{0}$ the original CMA-ES is recovered ($\boldsymbol{I}_k^{\sigma}$ in (6) becomes the identity).

Default parameters We set the usual default parameters $\mu = \left\lfloor \frac{\lambda}{2} \right\rfloor$, $w_i = \frac{\ln(\mu+1) - \ln i}{\sum_{j=1}^{\mu}(\ln(\mu+1) - \ln j)}$, $\mu_{\text{w}}^{-1} = \sum_{i=1}^{\mu} w_i^2$, and $d_\sigma = 1 + c_\sigma + 2\max\left(0, \sqrt{\frac{\mu_{\text{w}}-1}{n+1}} - 1\right)$ usually close to one, $\boldsymbol{C}_k^{-\frac{1}{2}}$ is symmetric and satisfies $\boldsymbol{C}_k^{-\frac{1}{2}} \boldsymbol{C}_k^{-\frac{1}{2}} = (\boldsymbol{C}_k)^{-1}$. The remaining learning parameters are chosen as $c_\sigma = \frac{\mu_{\text{w}}+2}{n+\mu_{\text{w}}+3}$, $c_{\text{c}} = \frac{4+\mu_{\text{w}}/n}{n+4+2\mu_{\text{w}}/n}$, $c_1 = \frac{2}{(n+1.3)^2+\mu_{\text{w}}}$ and $c_\mu = \min\left(1 - c_1,\, 2\frac{\mu_{\text{w}}-2+1/\mu_{\text{w}}}{(n+2)^2+\mu_{\text{w}}}\right)$.

3 Discussion

The implemented modifications come into play, when the original variation of a single component of the solution vector becomes smaller than its a priory given granularity. Asuming this granularity to be one, variations of ± 1 are introduced into some candidate solutions, with another small probability for an additional variation. The modification is a simple rescue scheme, because the variations are independent, that is, their covariances are zero. They facilitate mainly a simple local search in coordinate directions.

We discuss introduced parameters.

- The variance threshold for setting elements in the masking matrix $\boldsymbol{I}_k^{\text{R}}$ to one (therefore possibly applying an integer mutation) is $2\sigma_k \boldsymbol{C}_k^{\frac{1}{2}} < \boldsymbol{S}^{\text{int}}$. This threshold is sufficiently large, because variables without integer mutation are guaranteed to have a chance for mutation of at least 32% and a chance of at least 2% to mutate into the more unfavorable variable value.

- λ_k^{int} determines the number of candidate solutions that carry an integer mutation and depends on the number of variables that need to be mutated. Half of the λ new candidate solution do not have an integer mutation, such that when all integer variables are chosen correctly a reasonable convergence speed to the optimum can be achieved for the remaining continuous variables.

- The secondary integer mutation R'' is meant to be useful when more than one integer components must be changed or a value larger than one must be added/subtracted to find a better solution. The mutation rate is low and we suppose that the practical relevance of the secondary mutation is rather limited. If R'' is essential, a larger mutation rate for R'' might be desirable.

- The variance threshold for masking out variables in step-size control by setting diagonal elements of $\boldsymbol{I}_k^\sigma$ to zero is $5\,\sigma_k\,\boldsymbol{C}_k^{\frac{1}{2}}/\sqrt{c_\sigma} < \boldsymbol{S}^{\text{int}}$. The value $1/c_\sigma$ is the backward time horizon of the cumulation for step-size adaptation. The standard deviation of the normal random walk over $1/c_\sigma$ iterations with given σ and covariance matrix amounts to $\sigma\,\boldsymbol{C}^{\frac{1}{2}}/\sqrt{c_\sigma}$. In order to detect a selection effect, borders of the granularity must be passed and this standard deviation cannot be orders of magnitudes smaller than the granularity width.

  If all variables have positive granularity the step-size might not become small enough. For this reason even in this case λ_k^{int} is smaller than λ, which seems to be an effective, but most probably inferior solution to the problem.

- We have chosen to resample at least once the best integer setting from the last generation. The reason is that even if this setting happens to be good, with weighted recombination the new mean might still have a different setting. If the best setting was based on a comparatively seldom event resampling will help to retain it.

- When all variables undergo the integer procedure and satisfy the step-size masking condition, σ remains unchanged. This seems not necessarily justified and depending $\sigma_k^2\boldsymbol{C}_k$ increasing or decreasing might be appropriate.

Productive codes of CMA-ES support a number of termination criteria, see e.g. [1, 3]. With the new integer component, many of them need to be carefully reconsidered and, for example, only applied in non-integer components.

4 Simulations

We show a few experiments, first on the axis-parallel ellipsoid

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} 10^{6\frac{i-1}{n-1}} v_i^2 \tag{7}$$

with $\boldsymbol{v} = \boldsymbol{x}$. The variables with a high index have a high sensitivity. Usually, we expect integer variables to have a higher sensitivy, which is, fortunately also the easier case. We have tested a variety of different cases and show some of the more difficult ones in the following.

All experiments are done with initialization of $\boldsymbol{m}_{k=0} = \mathbf{1}$ and $\sigma_{k=0} = 10$. In Figure 1 the two typical outcomes are shown when variables 2, 5 and 8 are integer variables in dimension 10. Above, the algorithm does not approach the optimum satisfactorily, because the second variable retains a value far from zero. The variable is stuck on a plateau. Below, the algorithm approaches the optimum, which happens in about 20% of the cases. In Figure 2 the variables 1, 4 and 7 are integer variables. Variable 1 is the least sensitive variable and therefore the probability to approach the optimum is below 3%. If additionally variable 2 is an integer variable, the probability drops below 1%.

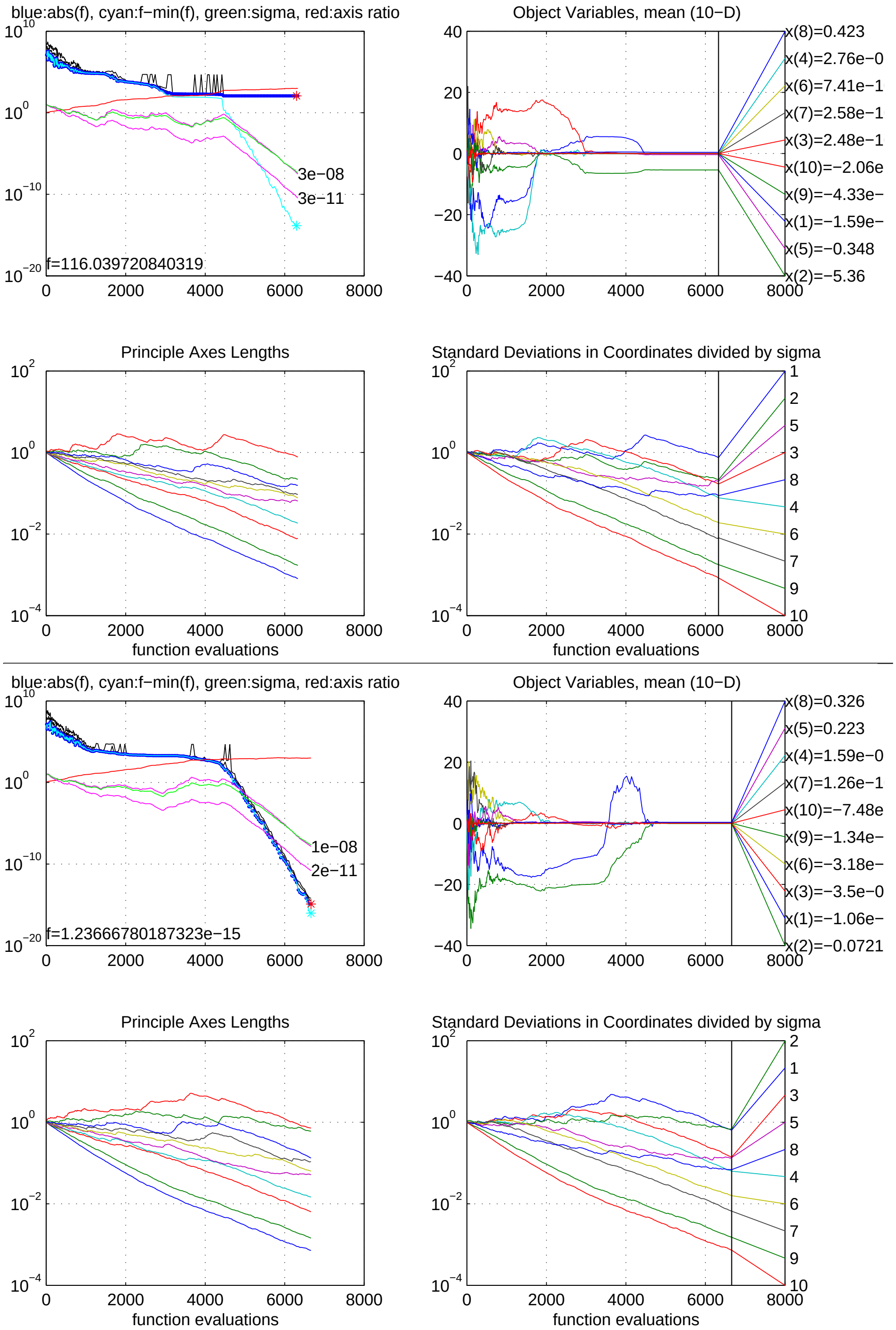

Figure 1: Two typical runs without integer handling on the axis-parallel ellipsoid, where variables 2,5,8 are integer variables. Below is the less common case (about 20%) where the optimum is well approached

Figure 3 shows the same situation with integer handling applied.

Increasing the population size has, at first sight, a similar effect as the integer handling. For population size $\lambda = 100$ the probability to approach the optimum well is high. It takes about twice as long to reach a function value of 10^{-10}. The more severe disadvantage might be that the result heavily depends on the initial step-size σ_0. With a small initial step-size the probability to converge is much smaller even with a large population size. This is particular true, if integer variables with a low variable index are prevalent.

On the rotated ellipsoid, the solution vector $\boldsymbol{x}$ undergoes a fixed random rotation according to [8, Fig.6] (resulting in $\boldsymbol{v}$) which is used in the RHS of (7). The topography becomes highly multimodal in this case. From the curves of the object variables in Figure 5 (below), we reckon the probability to approach the optimum being below $(1/n)^{1/n}$. With a population size 500 the problem can be solved in about 50% in about 50 000 function evaluations. Integer handling alone is not feasible to solve this problem, but we found some cases, where it significantly increases a small success probability.

5 Conclusion

We have introduced a comparatively simple rescue scheme that prevents CMA-ES to get stuck for a long time on single values of integer variables. The scheme is applicable to any discretized variable, however not very sensible in the binary case or for k-ary variables with, say, $k < 10$. We have shown simple test cases where the scheme raises the probability to find the optimal solution from less than 1% to 100%. The scheme disregards correlation between integer variables when variable variations become small and hence deserves to be referred to as a simple rescue scheme.

Acknowledgment This work has been partially founded by the FUI of System@tic Paris-Region ICT cluster through contract DGT 117 407 Complex Systems Design Lab (CSDL).

References

[1] A. Auger and N. Hansen. A restart CMA evolution strategy with increasing population size. In *Proc. IEEE Congress On Evolutionary Computation*, pages 1769–1776, 2005.

[2] N. Hansen. The CMA evolution strategy: a comparing review. In J.A. Lozano, P. Larranaga, I. Inza, and E. Bengoetxea, editors, *Towards a new evolutionary computation. Advances on estimation of distribution algorithms*, pages 75–102. Springer, 2006.

[3] N. Hansen. Benchmarking a BI-population CMA-ES on the BBOB-2009 function testbed. In *Workshop Proceedings of the GECCO Genetic and Evolutionary Computation Conference*, pages 2389–2395. ACM, July 2009.

[4] N. Hansen and S. Kern. Evaluating the CMA Evolution Strategy on multimodal test functions. In X. Yao et al., editors, *Parallel Problem Solving*

from Nature PPSN VIII, volume 3242 of *LNCS*, pages 282–291. Springer, 2004.

[5] N. Hansen, S. D. Müller, and P. Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation. *Evolutionary Computation*, 11(1):1–18, 2003.

[6] N. Hansen, S.P.N. Niederberger, L. Guzzella, and P. Koumoutsakos. A method for handling uncertainty in evolutionary optimization with an application to feedback control of combustion. *IEEE Transactions on Evolutionary Computation*, 13(1):180–197, 2009.

[7] N. Hansen and A. Ostermeier. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *ICEC96*, pages 312–317. IEEE Press, 1996.

[8] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

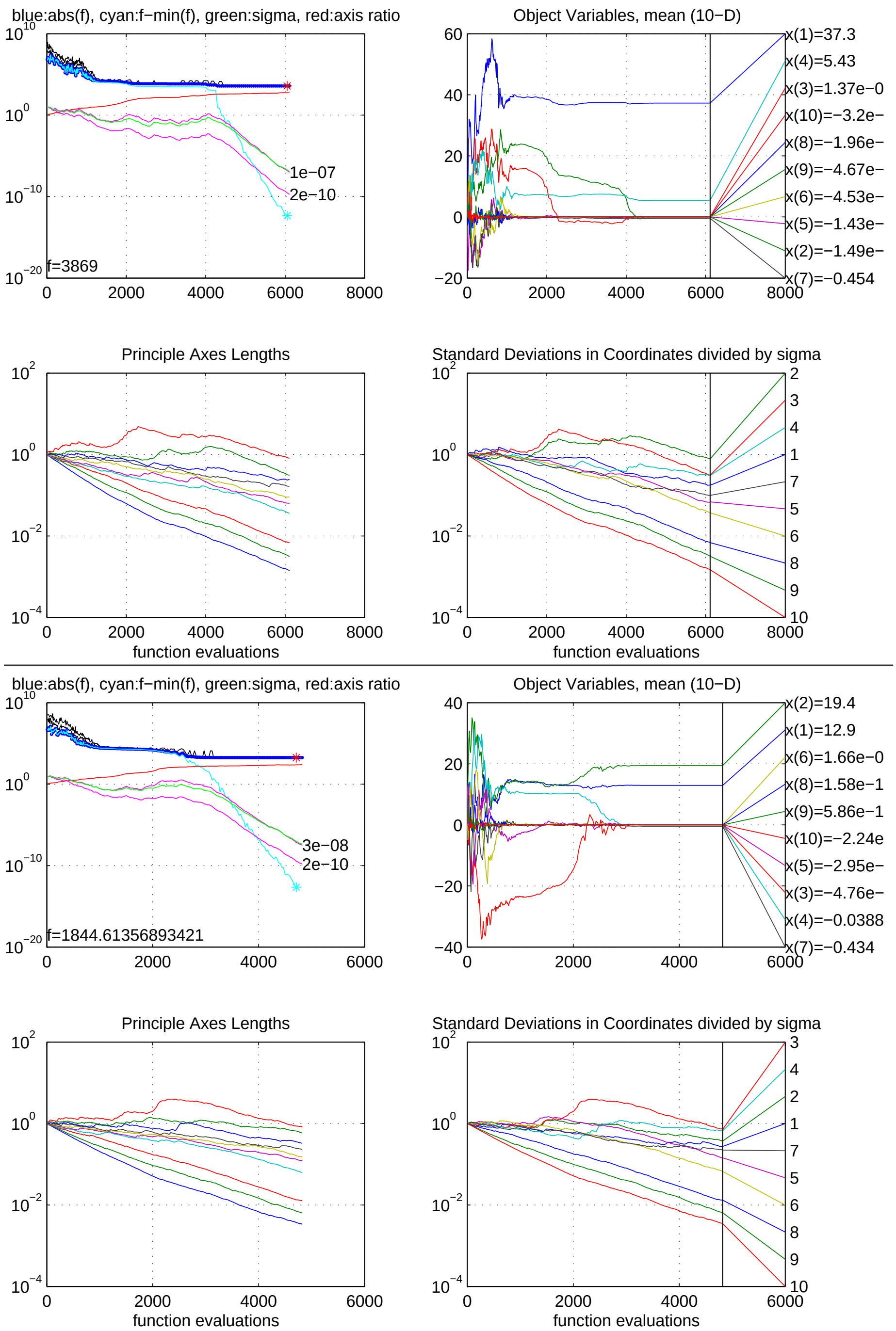

Figure 2: Two typical runs on the axis-parallel ellipsoid without integer handling, where variables 1,4,7 are integer variables (above) and additionally variable 2 (below). The optimization gets usually stuck. Only in about 3% the optimum is well approached in the above case. If also the 2nd variable is an integer variable the probability is $< 1\%$

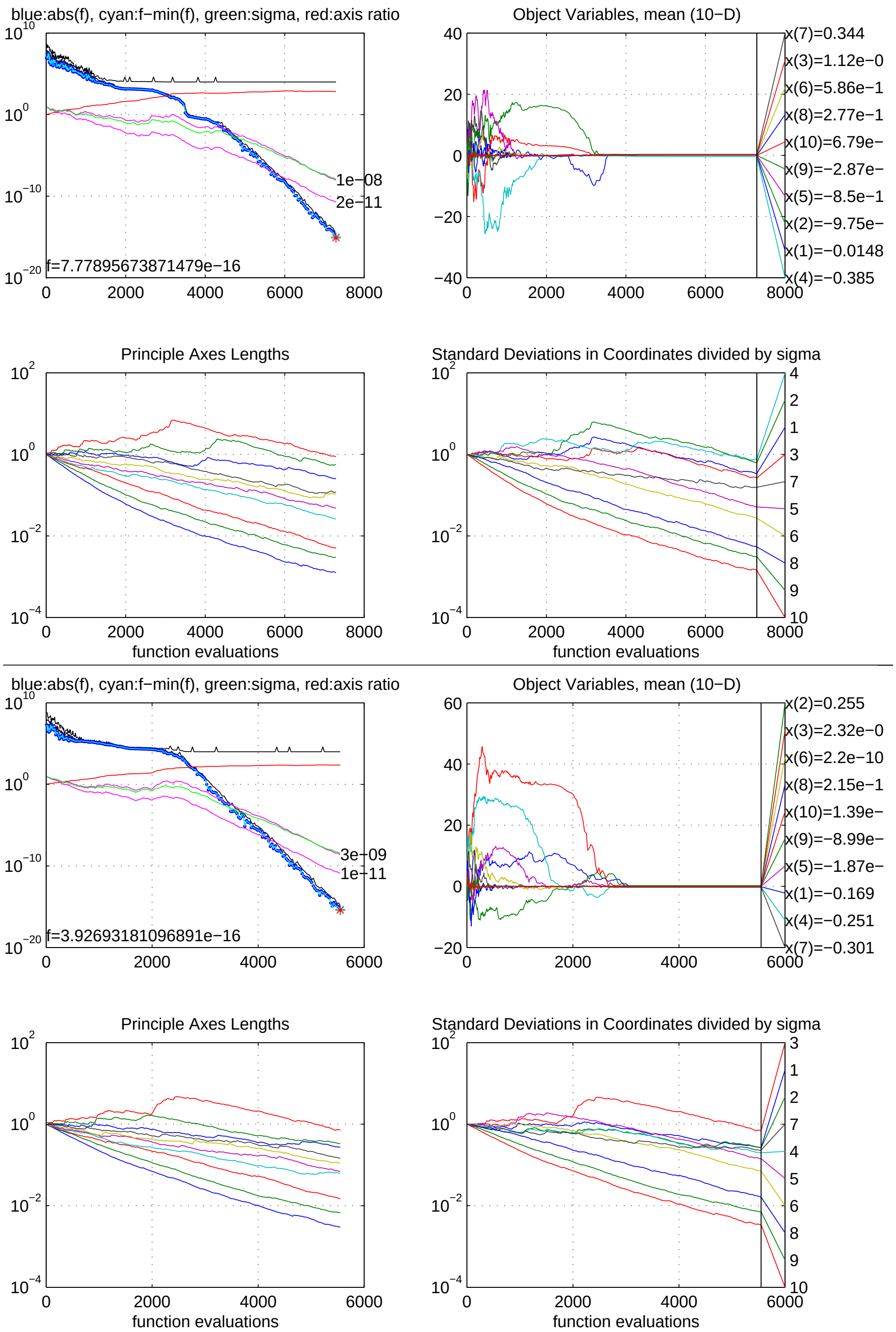

Figure 3: Two typical runs on the axis-parallel ellipsoid analogous to Figure 2 but with integer handling. The optimum is always approached well

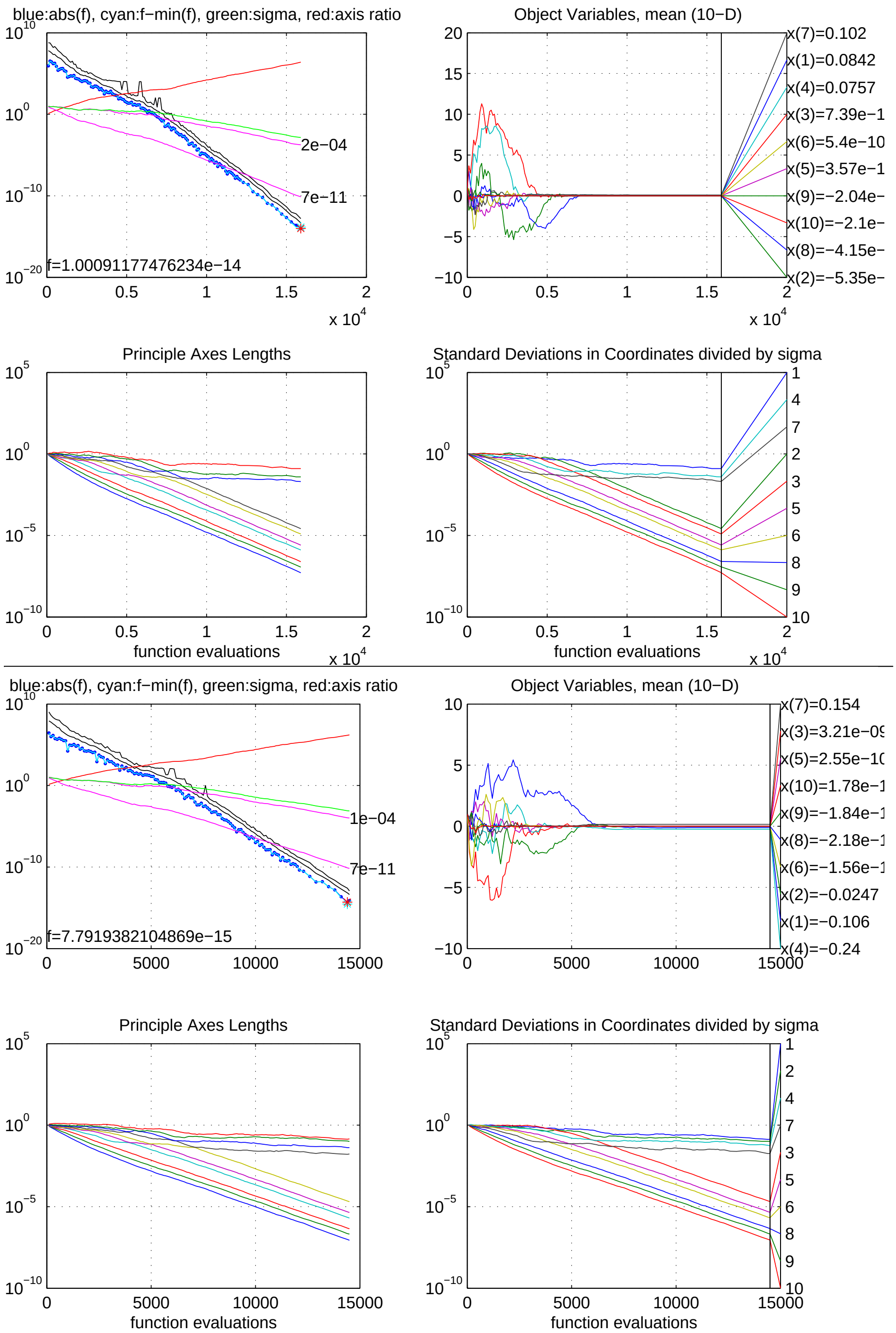

Figure 4: Two typical runs on the axis-parallel ellipsoid analogous to Figure 2 but with population size $\lambda = 100$. The optimum is usually approached well, but the performance depends essentially on the initial step-size

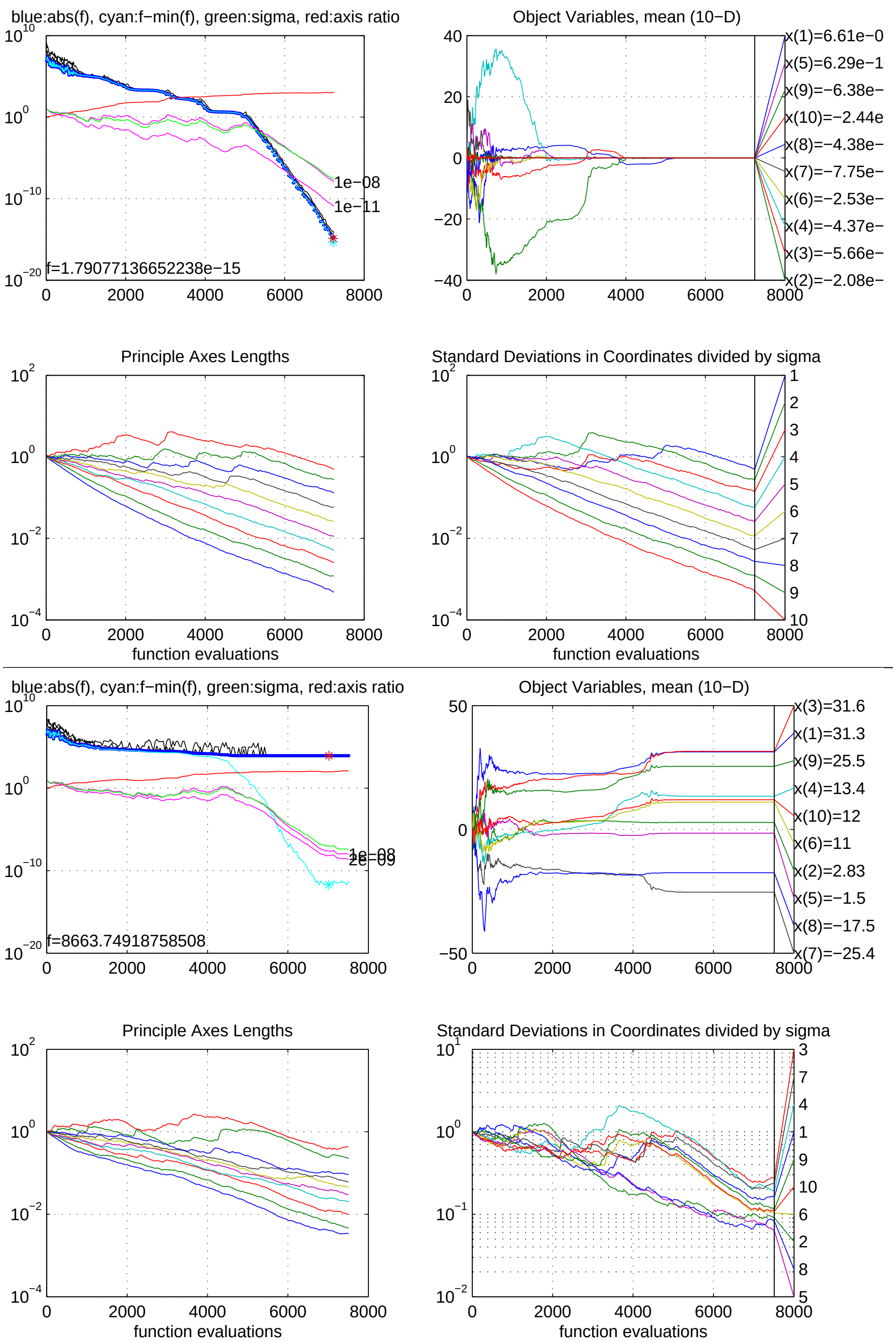

Figure 5: Two typical runs on the rotated ellipsoid. Above, all variables are continuous and the optimum is approached well. Below, variables 2,5,8 are integer variables and the optimum was never well approached (unless a much larger population size is used). Due to the integer variables, the rotated ellipsoid has a significantly different topography compared to the axis-parallel ellipsoid. The topography becomes highly multimodal

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399